A Validity Measure for a New Hybrid Data Clustering

Mahmut Hekim, Umut Orhan
Gaziosmanpasa University, Electronics and Computer Dept.
Tasliciftlik, 60250, Tokat, Turkey
mhekim@gop.edu.tr, umutorhan@gop.edu.tr

Abstract

Data clustering method is a process of putting similar data into groups. A clustering method partitions a data set into several groups such that the similarity within a group is larger than among groups. It has been playing an important role in solving many problems in image processing and pattern recognition. In this paper, a new method called hybrid clustering method is obtained by the most representative clustering methods, and considered by a new validity measure. We define Y as a clustering validity function which measures to selecting optimal number of clusters using the measurement value of Y, which is coverage area. It is compared with conventional validity functions, partition coefficient PC and compactness and separation validity function G in several data sets.

1. Introduction

Data clustering is grouping of objects into homogenous groups based on same object features; and it is considered an interesting approach for finding similarities in data and putting similar data into groups. This approach is an important for image processing; remote sensing, data mining, and pattern recognition. Generally speaking, clustering is one method to find most similar groups from given data, which means that data belonging to one cluster are the most similar; and data belonging to different cluster are the most dissimilar. In the literature, researchers have proposed many solutions for this issue based on different theories, and many surveys focused on special types of clustering algorithm have been presented [1], [2], [3], [4], [5].

Clustering algorithms are used not only to categorize data, but are also useful for data compression and model construction. By finding similarities, similar data can be represented with fewer symbols. Also, if we can find groups of data, we can build a solution based on those groupings.

In this paper, the most representative clustering techniques are reviewed; and by improving of existing clustering techniques and combining by hard clustering and subtractive clustering, we propose a new hybrid approach. This method is called "hybrid clustering method" and it is the subject of this paper.

The number of clusters in data is related to cluster validity problem which is how well it has identify the structure that is present in the data. Several validity functions such as partition coefficient [6], classification entropy [6], proportion exponent [6], [7], csc index [8] and so on, have been used for measuring validity mathematically. The common approach to find optimal number of clusters in data is to record the value of validity functional as a function of c and choose as optimum the number of clusters for which large change occurs in the values. Therefore, it requires human interpretation and subjective analysis of what is to be considered a large change in the values. For improving hybrid clustering method, we define Y as validity function which measures optimal coverage area of clusters and propose a new fuzzification method which is the independent on human interpretation by using Y function.

2. Data Clustering Overview

In this section, a detailed discussion of each technique is investigated. Hard clustering, which is also called K-means clustering, is an algorithm based on finding data clusters in a data set such that the cost function of dissimilarity measure minimized. In most cases this dissimilarity measure is chosen as the Euclidean distance [6], [9].

A set of n vectors x_j are to be partitioned into c groups G_i (where i=1,…,c and j=1,…,n). The cost function, based on the Euclidean distance between a vector x_k in group j and the corresponding cluster center c_i, can be defined by:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \left\| x_k - c_i \right\|^2 \right) \quad (1)$$

where J_i is the cost function within group *i*. The partitioned groups are defined by a $c \times n$ binary membership matrix *U*, where the element u_{ij} is 1 if the *j*th data point x_j belongs to group *i*, and 0 otherwise. Once the cluster centers c_i are fixed, the minimizing u_{ij} for (Eq. 1) can be derived as follows:

$$u_{ij} = \begin{cases} 1 \text{ if } \left\| x_j - c_i \right\|^2 \le \left\| x_j - c_k \right\|^2, \ \forall k \ne i \\ 0 \text{ otherwise} \end{cases} \quad (2)$$

This means that x_j belongs to group i if c_i is the closest center among all centers. On the other hand, if the membership matrix is fixed, i.e. if u_{ij} is fixed, then the optimal center c_i that minimizes (Eq. 1) is the mean of all vectors in group *i*:

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \quad (3)$$

where $|G_i|$ is the size of G_i , or $|G_i| = \sum_{j=1}^{n} u_{ij}$.

The method determines the cluster centers c_i and the membership matrix *U* iteratively. The performance of the K-means algorithm depends on the initial positions of the cluster centers, thus it is advisable to run the algorithm several times, each with a different set of initial cluster centers [3].

Fuzzy C-means clustering relies on the basic idea of K-means clustering. But, each data point belongs to a cluster with a degree of membership while in K-means every data point either belongs to a certain cluster or not. So fuzzy clustering employs fuzzy partitioning such that a given data point can belong to several groups with the degrees of membership between 0 and 1. However, Fuzzy clustering also uses a cost function. In this method, the membership matrix *U* is allowed to have elements with values between 0 and 1. Thus the summation of degrees of membership of a data point to all clusters is always equal to 1 [3], [6], [7]:

$$\sum_{i=1}^{c} u_{ij} = 1, \quad \forall j = 1, ..., n \quad (4)$$

The cost function for FCM is a generalization of (Eq. 1):

$$J(u, c_1, ..., c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \quad (5)$$

where u_{ij} is between 0 and 1; c_i is the cluster center of fuzzy group *i*; $d_{ij} = \left\| c_i - x_j \right\|$ is the Euclidean distance between the *i*th cluster center and the *j*th data point; and $m \in [1, \infty]$ is a weighting exponent. The necessary conditions for (Eq. 5) to reach its minimum are

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \quad (6)$$

and

$$u_{ji} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \quad (7)$$

The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. As in HCM, the performance of FCM depends on the initial membership matrix values. Thereby it is advisable to run the algorithm for several times with different values of membership degrees of data points [3].

Subtractive clustering method is based on a measure of the density of data points in the feature space [1]. The idea is to find regions in the feature space with high densities of data points. The point with the highest number of neighbours is selected as center for a cluster. Since each data point x_i is a candidate for cluster center, a density measure for first cluster center $c1$ is defined as

$$D_i = \sum_{j=1}^{n} \exp\left(- \frac{\left\| x_i - x_j \right\|^2}{(r_a / 2)^2} \right) \quad (8)$$

where r_a is a positive constant representing a neighbourhood radius. Hence, a data point will have a high density value if it has many neighbouring data points. The first cluster center x_{c1} is selected as the point having the largest density value D_{c1}. Next, the density measure of each data point x_i is revised as follows:

$$D_i = D_i - D_{c_1} \exp\left(- \frac{\left\| x_i - x_{c_1} \right\|^2}{(r_b / 2)^2} \right) \quad (9)$$

where r_b is a positive constant which defines a neighbourhood having measurable reductions in density measure. Therefore, the data points near the first cluster center x_{c1} will have significantly reduced density measure. After revising the density function given by (Eq. 9), the next cluster center is chosen as the point having the greatest density value. This process continues until a sufficient number of clusters are attained [6].

The cluster validity functions which have been used as a measure of the quality of clustering are as follows [6], [10].

- Partition Coefficient (PC):

$$PC(U, c) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^2}{n}; \quad \max_c \left[\max_U \{ PC(U; c) \} \right] \quad (10)$$

- Compactness and separation validity function (G):

$$C = \frac{2}{n-1}\sum_{j_1=1}^{n-1}\sum_{j_2=j_1+1}^{n}\sum_{i=1}^{c} d^2(X_{j1}, X_{j2}) \qquad (11)$$

$$D = \frac{1}{n^2}\sum_{j1=1}^{n}\sum_{j2=1}^{n} d^2(X_{j1}, X_{j2}) \qquad (12)$$

$$G = \frac{D}{C} \qquad (13)$$

The compactness and separation validity function is defined as the ratio of separation index D to the compactness index C, i.e., $G = D/C$.

3. Hybrid Clustering Method

In general clustering techniques, according to the dimension of data and the number of clusters, desired result may not be reached for many applications. In hard clustering, if the number of clusters is sufficient, then this algorithm can obtain optimum result, otherwise can not. Thus, optimum result is related to the algorithm's truth, directly. Cluster centers detected by algorithms are fault when the number of clusters is given incomplete or excess.

Hybrid clustering method offered by this study is a new approach to finding the number of clusters and allowing being reduced computation times, significantly. This new approach uses both of K-means and Subtractive clustering methods. First, it is started by using Subtractive clustering to detect the initial cluster center points for K-means. Next, K-means method detects certain cluster centers by using them.

Hybrid clustering method also uses the cost function. However, the problem of detecting optimum number of clusters in the others is removed by this new approach. While the performance of common clustering methods depends on the initial membership matrix values and the number of clusters; it removes this problem. Thus, we propose an approach to select optimal number of clusters. When n_c is equal to 1 it is used (Eq. 8) and otherwise (Eq. 9), where n_c is the number of iterations used for hybrid clustering; and when $c_i = c_{i-1}$ is provided then the computation is stopped and done $n_c = i - 1$. The final of obtained value is the number of clusters.

The cost function J and the membership matrix U are computed according to (Eq. 1 and 2), respectively. If the cost function J is below a certain tolerance value, iteration is stopped; and the cluster centers are updated according to (Eq. 3).

Distances among cluster centers are computed then minimum (R_1) and maximum (R_2) values detected by using Euclidean distance, and R is the coverage area described in (Eq. 13). Each data point belongs to all cluster centers with the ratio of distances from each cluster centers. Thus, a new membership matrix U is calculated by using (Eq. 10). But, it is supposed that a data point which is more distant from any cluster center than R_2 value is affected by a lost cluster center. Therefore, R_2 must be taken account of (Eq. 14 and Eq. 15).

$$R = \frac{R_1}{R_2} \qquad (14)$$

$$u_{ij} = \frac{S_{\min}}{S_j}\sum_{j=1}^{c}\sum_{i=1}^{ks}\left(\frac{\sum_{i=1}^{ks}\frac{1}{d_{ij}^2}}{d_{ij}^2}\right), d_{ij} = \left\|c_i - x_j\right\|^2 \qquad (15)$$

where $S_j = \sum_{i=1}^{ks}\left\|c_i - x_j\right\|$ and $S_{\min} = \min\{S_j\}$.

This approach determines the cluster centers c_i and the membership matrix U using the following steps:

*Step*1. Normalize the data points.
*Step*2. Calculate the density measure of each data point according to (Eq. 8 and 9).
*Step*3. Select as n_cth cluster center the point having the largest density value, and continue until the number of clusters n_c is detected.
*Step*4. Initialize the cluster centers with computed in Step 3.
*Step*5. Compute the hard membership matrix U using by (Eq. 2).
*Step*6. Calculate the cost function according to (Eq. 1). Go to step 8 if it is below a threshold value.
*Step*7. Update the cluster centers according to (Eq. 3). Go to Step 5.
*Step*8. Determine the fuzzy membership matrix U using by (Eq. 15).

4. Numeric Examples

To show the validity of the proposed method, we prepare data sets, iris data and blood pressure data. For this experiment, we examine K-means, subtractive clustering and hybrid clustering methods by adding validity measures. We compare our approach with validity strategies which find optimal number of clusters using *PC*, *G* and *R* defined as (Eq. 10, 11 and 14), respectively.

Figure 1 shows subtractive clustering method for different number of clustering. Whereas, Figure 4 shows hybrid clustering techniques finding optimum number of clusters and detecting cluster centers without very long calculation times.

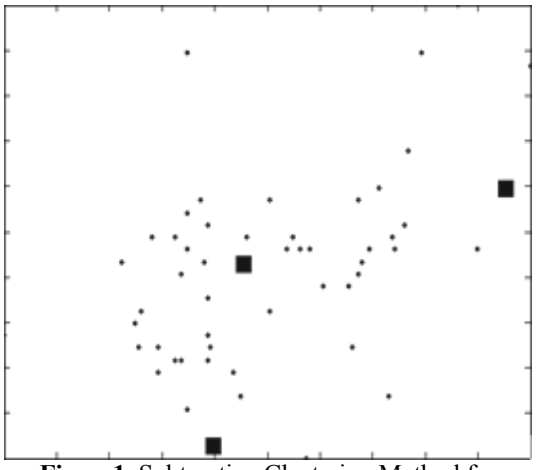

Figure1. Subtractive Clustering Method for blood pressure data set

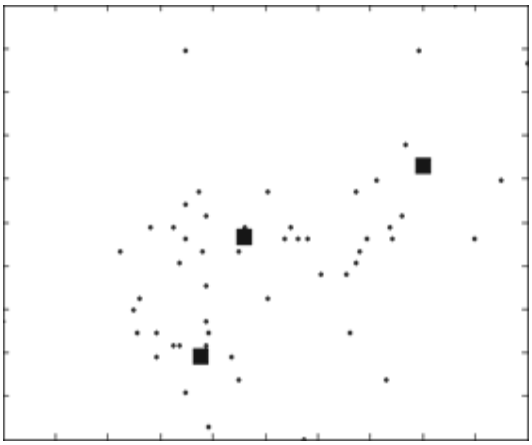

Figure 2. Hybrid Clustering Method for blood pressure data set

According to the result of numeric examples, it is clear from the results that choosing n_c small or big are resulted in poor accuracy, whereas hybrid approach offered in this paper results the highest accuracy.

Example 1. Anderson Iris Data set [10] consists of 150 four dimensional vectors. This data set has often been used as a standard for testing clustering algorithms and validity function [6], [11]. Result for this data set is summarized in Table 2, and it shows the optimal number of clusters using *PC* is 2 whereas using *PC* and *R* is 3, which is equal to the already known optimal number. When the number of clusters is 2, R_1 is equal to R_2 , so this situation is not taken into account.

Table 1. Measurement values for the iris data

C	*PC*	*G*	*R*
c=4	0.46	1.42	0.22
c=3	0.57	1.78 *	0.43 *
c=2	0.69 *	1.69	-

Example 2. Blood pressure data set [10] consists of 53 two dimensional vectors that have three clusters. This data set has often been used as a standard for testing clustering algorithms and validity function [6], [11]. Result for this data set is summarized in Table 3. It shows the optimal number of clusters using *G* is 2, whereas using *PC* and *R* is 3 which is equal to the already known optimal number.

Table 2. Measurement values for the blood pressure data

C	PC	G	R
c=6	0.33	0.84	0.26
c=4	0.43	0.95 *	0.33
c=3	0.47 *	0.84	0.47 *

Table 3 shows the cluster centers obtained by hard clustering, fuzzy clustering, subtractive clustering techniques and a new hybrid approach.

Table 3. The cluster centers obtained by those methods

	Hard Clustering		**Fuzzy Clustering**	
	Log of dose	Blood pressure	Log of dose	Blood pressure
*Cluster*1	2.46	74.33	2.36	69
*Cluster*2	1.71	66	2.04	68
*Cluster*3	2.05	61.44	1.72	59
	Subtractive Clust.		**Hybrid Clustering**	
	Log of dose	Blood pressure	Log of dose	Blood pressure
*Cluster*1	1.9	67	1.91	68.43
*Cluster*2	2.7	73	2.46	74.33
*Cluster*3	1.81	52	1.78	58.33

5. Conclusion

The issue of validity for fuzzy clustering has been neglected with few notable exceptions. However, if cluster analysis is to make a significant contribution to engineering applications, much more attention must be paid to fundamental questions of optimal number of clusters. Hybrid clustering method offered by this study is a new approach to finding the number of clusters and allowing being reduced computation times, significantly. This new approach uses both of K-means and Subtractive clustering methods.

In this paper, we have defined *R* as a clustering validity criterion. It measures the overall coverage area of all clusters. Moreover we have proposed a new approach to selecting optimal number of clusters combined with K-means and subtractive clustering methods. We have tested this approach on the two data sets. Compared with the validity

strategies using *PC* and *G*, we have noted that the proposed approach produce more valid result.

6. References

[1] J. S. R. Jang, C. T. Sun, and E. Mizutani, *Neuro-Fuzzy and Soft Computing – a Computational Approach to Learning and Machine Intellence*, Prentice Hall, 1997.

[2] J. Yu, "General C-Means Clustering Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.27, No.8, pp.1197-1211, August 2005.

[3] K. Hammouda, and F. Karay, A Comparative Study of Data Clustering Techniques, Course Project, 2000.

[4] A. Baraldi, and P. Blonda, "A Survey of Fuzzy Clustering Algorithms for Pattern Recognetion-Part I and II," *IEEE Trans. Systems, Man, and Cybernetics, Part B*, vol. 29, no. 6, pp. 778–801, 1999.

[5] J. Han and M. Kamber, *Data mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2000.

[6] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, *Plenum Press*, NewYork,1981.

[7] T.J. Ross, *Fuzzy Logic with Engineering Applications*, McGraw-Hil., 1995.

[8] E. Xei. and G. Beni, "A Validity Measure for Fuzzy Clustering", *IEEE Trans. on Pattern Analysis Machine Intelligence*, vol. PAMI-13, no. 8, 1991.

[9] S. Chopra, R. Mitra and V. Kumar, "Identification of Rules Using Subtractive Clustering with Application to Fuzzy Controllers," *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, Shanghai, 26-29 August 2004, pp. 4125-4130.

[10] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski, *A Handbook of Small Data Sets*, Chapman and Hall, London, 1994.

[11] Y. T. Chein, *Interactive Pattern Recognition*. Marcel Dekker, NewYork, 1978.